

# Controlling for Unobserved Confounds in Classification Using Correlational Constraints

Virgile Landeiro and Aron Culotta

Department of Computer Science  
Illinois Institute of Technology  
Chicago, IL 60616  
vlandeir@hawk.iit.edu, aculotta@iit.edu

## Abstract

As statistical classifiers become integrated into real-world applications, it is important to consider not only their accuracy but also their robustness to changes in the data distribution. In this paper, we consider the case where there is an unobserved confounding variable  $z$  that influences both the features  $\mathbf{x}$  and the class variable  $y$ . When the influence of  $z$  changes from training to testing data, we find that the classifier accuracy can degrade rapidly. In our approach, we assume that we can predict the value of  $z$  at training time with some error. The prediction for  $z$  is then fed to Pearl’s back-door adjustment to build our model. Because of the attenuation bias caused by measurement error in  $z$ , standard approaches to controlling for  $z$  are ineffective. In response, we propose a method to properly control for the influence of  $z$  by first estimating its relationship with the class variable  $y$ , then updating predictions for  $z$  to match that estimated relationship. By adjusting the influence of  $z$ , we show that we can build a model that exceeds competing baselines on accuracy as well as on robustness over a range of confounding relationships.

## 1 Introduction

Statistical classifiers have become widely used to inform important decisions such as whether to approve a loan (Hand and Henley 1997), hire a job candidate (Miller 2015), or release a criminal defendant on bond (Monahan and Skeem 2016). Given the significant real-world consequences of such decisions, it is critical that we can identify and remove sources of systematic bias in classification algorithms. For example, some evidence suggests that existing criminal recidivism models may be racially biased (Angwin et al. 2016).

One important type of classifier bias arises from *confounding variables*. A confounder  $z$  is a variable that is correlated both with the input variables (or features)  $\mathbf{x}$  and the target variable (or label)  $y$  of a classifier. When  $z$  is not included in the model, the true relationship between  $\mathbf{x}$  and  $y$  can be improperly estimated; in the social sciences – originally in econometrics – this is called omitted variable bias. While omitted variable bias is a core focus of social science (King, Keohane, and Verba 1994), it has received much less attention in machine learning communities, where prediction accuracy is the main concern. Con-

founding variables can be particularly problematic in high-dimensional settings, such as text classification, where models may contain thousands or millions of parameters, making manual inspection of models impractical. The common use of text classification in computational social science applications (Lazer et al. 2009) further adds to the urgency of the problem.

Several studies with interests in public health focused on tracking the influenza rates in the USA by using Twitter as a sensor (Paul and Dredze 2011). These studies demonstrated that machine learning offers more accurate, inexpensive, and fast tracking methods than what is currently used by the CDC. De Choudhury, Counts, and Horvitz built models to predict postpartum changes in emotion and behavior using Twitter data and managed to identify mothers who will change significantly following childbirth with an accuracy of 71% using observations about their prenatal behavior (De Choudhury, Counts, and Horvitz 2013). In a more recent study, Koratana et al. collected Yik Yak data – an anonymous social network popular among students – to study anonymous health issues and substance use on college campuses (Koratana et al. 2016). The results of these studies are encouraging for the field of computational social science but only a few of them are taking into account the effect of possible confounders. A growing body of work tries to mitigate the effect of observed confounding variables using causal inference techniques. For instance, Cunha, Weber, and Pappa use a matching approach for causal inference to estimate the effect of online support on weight loss using data from Reddit, and De Choudhury et al. leverage propensity score matching to detect users that transition from posting about mental health concerns to posting about suicidal ideation on Reddit. In this paper, we wish to provide methods for researchers in computational social sciences to conduct observational studies while controlling for confounding variables even though these might not be directly observed.

In recent work (Landeiro and Culotta 2016), a text classification algorithm was proposed based on Pearl’s back-door adjustment (Pearl 2003) as a framework for prediction that controls for an observed confounding variable. It was found that this approach results in classifiers that are significantly more robust to shifts in the relationship between confounder  $z$  and class label  $y$ . However, an important limitation of this prior work is that it assumes that a training set is available in

which every instance is annotated for both class label  $y$  and confounder  $z$ . This is problematic because there are many confounders we may want to control for (e.g., income, age, gender, race/ethnicity) that are often rarely available and difficult for humans to label, particularly in addition to the primary label  $y$ .

A natural solution is to build statistical classifiers for confounders  $z$ , and use the predicted values of  $z$  to control for these confounders. However, the measurement error of  $z$  introduces *attenuation bias* (Chesher 1991) in the back-door adjustment, resulting in classifiers that are still confounded by  $z$ .

In this paper, we present a classification algorithm based on Pearl’s back-door adjustment to control for an *unobserved* confounding variable. Our approach assumes we have a preliminary classifier that can predict the value of the confounder  $z$ , and that we have an estimate of the error rate of this  $z$ -classifier. We offer two methods to adjust for the mislabeled  $z$  to improve the effectiveness of back-door adjustment. A straightforward approach is to remove training instances for which the confidence of the predicted label for  $z$  is too low. While we do find this approach can reduce attenuation bias, it must discard many training examples, degrading the  $y$ -classifier. Our second approach instead uses the error rate of the  $z$ -classifier to estimate the correlation between  $y$  and  $z$  in the training set. The assignment to  $z$  is then optimized to match this estimated correlation, while also maximizing classification accuracy. We compare our methods on two real-world text classification tasks: predicting the location of a Twitter user and predicting if a Twitter user is smoking or not. Both prediction tasks are using users’ tweets as input data and are confounded by gender. The resulting model exhibits significant improvements in both accuracy and robustness, with some settings producing similar results as fully-observed back-door adjustment.

## 2 Related Work

In the machine learning field, selection bias has received some attention (Zadrozny 2004; Bareinboim, Tian, and Pearl 2014). It arises when the population of a study is not selected randomly. Instead, some users are more inclined to be selected for the study than others, making it more difficult to draw conclusions from the general population. If we denote  $S$  whether or not an element of the population is selected, there is presence of selection bias when  $p(S = 1|X, Y) \neq p(S = 1)$ . Dataset shift (Quionero-Candela et al. 2009) is a similar issue that appears when the joint distribution of features and labels changes between the training dataset and the testing dataset (i.e.  $p_{tr}(X, Y) \neq p_{te}(X, Y)$ ). Covariate shift (Bickel, Brückner, and Schepfer 2009; Sugiyama, Krauledat, and Müller 2007) is a specific case of dataset shift in which only the inputs distribution is different from training to testing (i.e.  $p_{tr}(X) \neq p_{te}(X)$ ). Similarly, when the underlying target distribution  $p(Y)$  changes over time, either in a sudden way or gradually, then this is called concept drift (Tsybmal 2004; Widmer and Kubat 1996). Recent work has studied “fairness” in machine learning (Zemel et al. 2013; Hajian and

Domingo-Ferrer 2013) as well as attempted to remove features that introduce bias (Pedreshi, Ruggieri, and Turini 2008; Fukuchi, Sakuma, and Kamishima 2013). Kuroki and Pearl (2014) propose an extension of back-door adjustment to deal with measurement error in the confounder, but it does not scale well when  $\mathbf{x}$  is high dimensional, as in our setting of text classification.

Although all these types of biases are important to conduct a valid observational study, in this paper we direct our attention to the problem of learning under confounding bias shift. In other words, we aim to build a classifier that is robust to changes in the relation between the target variable  $Y$  of a classifier and an external confounding variable  $Z$ . Landeiro and Culotta (2016) use back-door adjustment for text classification, but assume confounders are observed at training time. This paper introduces methods to enable back-door adjustment to work effectively when confounders are unobserved and when the features are high dimensional.

## 3 Methods

In this section, we first review prior work using back-door adjustment to control for observed confounders in text classification. We then introduce two methods for applying back-door adjustments when the confounder is unobserved at training time and must instead be predicted by a separate classifier.

### 3.1 Adjusting for observed confounders

Suppose one wishes to estimate the causal effect of a variable  $\mathbf{x}$  on a variable  $y$  when a randomized controlled trial is not possible. If a sufficient set of confounding variables  $z$  is available, one can use the back-door adjustment equation as follows:

$$p(y|do(\mathbf{x})) = \sum_z p(y|\mathbf{x}, z) \times p(z) \quad (1)$$

The *back-door criterion* (Pearl 2003) is a graphical test that determines whether  $z$  is a sufficient set of variables to estimate the causal effect. This criterion requires that no node in  $z$  is a descendant of  $\mathbf{x}$  and that  $z$  blocks every path between  $\mathbf{x}$  and  $y$  that contains an arrow pointing to  $\mathbf{x}$ . Notice  $p(y|\mathbf{x}) \neq p(y|do(\mathbf{x}))$ : this  $do$ -notation is used in causal inference to indicate that an intervention has been made on  $\mathbf{x}$ . Omitting the predicted confounder  $z'$ , it depicts a standard discriminative approach to classification, e.g., modeling  $p(y|\mathbf{x})$  with a logistic regression classifier conditioned on the observed term vector  $\mathbf{x}$ . We assume that the confounder  $z'$  influences both the term vector through  $p(\mathbf{x}|z)$  as well as the target label through  $p(y|z')$ . The structure of this model ensures that  $z'$  meets the back-door criterion for adjustment.

Back-door adjustment was originally introduced for causal inference problems — i.e., to estimate the causal effect of performing action  $\mathbf{x}$  on outcome  $y$ . Recently, Landeiro and Culotta (2016) have shown that back-door adjustment can also be used to improve classification robustness. By controlling for a confounder  $z$ , the resulting classifier is robust to changes in the relationship between  $z$  and  $y$ .

From the perspective of standard supervised classification, the approach works as follows: Assume we are given a training set  $D = \{(\mathbf{x}_i, y_i)\}$ . If we suspect that a classifier trained on  $D$  is confounded by some additional variable  $z$ , we augment the training set by including  $z$  as a feature for each instance:  $D' = \{(\mathbf{x}_i, y_i, z_i)\}$ . We then fit a classifier on  $D'$ , and at testing time apply Equation 1 to classify new examples —  $p(y|\mathbf{x}) = \sum_z p(y|\mathbf{x}, z)p(z)$  — where  $p(z)$  is simply computed from the observed frequencies of  $z$  in  $D'$ . By controlling for the effect of  $z$ , the resulting classifier is robust to the case where  $p(y|z)$  changes from training to testing data.

In the experiments below, we consider the problem of predicting a user’s location  $y$  based on the text of their tweets  $\mathbf{x}$ , confounded by the user’s gender  $z$ . That is, in the training data, there exists a correlation between gender and location, but we want the classifier to ignore that correlation. When the above procedure is applied to a logistic regression classifier, the result is that the magnitudes of coefficients for terms that correlate with gender are greatly reduced, thereby minimizing the effect of gender on the classifier’s predictions.

### 3.2 Adjusting for unobserved confounders

In the previous approach, it was assumed that we had access to a training set  $D = \{(\mathbf{x}_i, y_i, z_i)\}$ ; that is, each instance is annotated both for the label  $y$  and confounder  $z$ . This is a burdensome assumption, given that ultimately we will need to control for many possible confounders (e.g., gender, race/ethnicity, age, etc.). Because many of these confounders are unobserved and/or difficult to obtain, it is necessary to develop adjustment methods that can handle noise in the assignment to  $z$  in the training data.

Our approach assumes we have an (imperfect) classifier for  $z$ , trained on a secondary training set  $D_z = \{(\mathbf{x}_i, z_i)\}$  — we call this the *preliminary study*, with the resulting *preliminary classifier*  $p(z|\mathbf{x})$ . This is combined with the dataset  $D_y = \{(\mathbf{x}_i, y_i)\}$ , used to train the primary classifier  $p(y|\mathbf{x})$ . The advantage of allowing for separate training sets  $D_y$  and  $D_z$  is that it is often easier to annotate  $z$  variables for some users than others; for example, Pennacchiotti and Popescu (2011) build training data for ethnicity classification by searching for online users that explicitly state their ethnicity in their user profiles.

After training on  $D_z$ , the preliminary classifier is applied to  $D_y$  to augment it with predicted annotations for confounder  $z$ :  $D = \{(\mathbf{x}_i, y_i, z'_i)\}_{i=1}^n$ , where  $z'_i$  denotes the predicted value of  $z_i$ . A tempting approach is to simply apply back-door adjustment as usual to this dataset, ignoring the noise introduced by  $z'$ . However, the resulting classifier will no longer properly control for the confounder  $z$  for at least two related reasons:

1. The observed correlation between  $y$  and  $z'$  in the training data will underestimate the actual correlation (i.e.,  $|r(y, z')| < |r(y, z)|$ ). This *attenuation bias* reduces the coefficients for the  $z$  features, which in turn prevents back-door adjustment from reducing the coefficients of features in  $\mathbf{x}$  that correlate with  $z$ .
2. Similarly, because some training instances have misla-

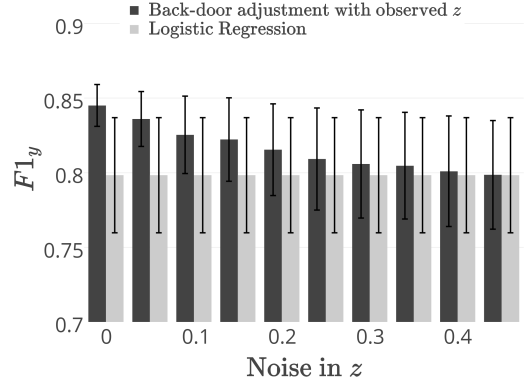


Figure 1: As measurement error in confounder  $z$  increases, the effectiveness of back-door adjustment decreases.

Noise	0.00	0.05	0.10	0.15	0.20
F1 std dev	0.028	0.037	0.052	0.056	0.062

Table 1: Evolution of the standard deviation of F1 scores in back-door adjustment for a given noise in  $z$ . The lower the standard deviation, the more robust the model.

beled annotations for  $z$ , it is more difficult to detect which features in  $\mathbf{x}$  correlate with  $z$ , thereby preventing back-door adjustment from reducing those coefficients.

To verify this claim, we conduct an experiment in which we observe  $z$  but we inject increasing amounts of noise in  $z$  (e.g., with probability  $p$ , change the assignment to  $z_i$  to be incorrect). In other words, we synthetically decrease the quality of our observations of  $z$  and we observe how that influences the performance of back-door adjustment. We then measure how the accuracy of the primary classifier for  $y$  varies on a testing set in which the influence of  $z$  is decreased (i.e.,  $z$  correlates strongly with  $y$  in the training set, but only weakly in the testing set). These experiments will be discussed in more detail in Section 4.

We can see in Figure 1 that the F1 score quickly decreases as we add more noise to the confounding variable annotations, indicating the need for new methods to adjust for unobserved confounders. Notice that when noise is 0, back-door adjustment greatly improves F1 (from .79 F1 with no adjustment to .85 F1), demonstrating the effectiveness of this approach when the confounder is observed at training time. In the following two sections, we propose two methods to fix these issues.

**Thresholding on confidence of  $z$  predictions** Our first approach is fairly simple; its objective is to directly reduce the number of mislabeled annotations in  $z'$ . Our preliminary model produces the value  $z'_i$  (the prediction of the true confounder  $z_i$ ) as well as  $p(z_i = z'_i|\mathbf{x}_i)$  (the confidence of the prediction; i.e., the posterior distribution over  $z$ ). We use these posteriors to remove predictions with low confidence. By setting a threshold  $\epsilon \in [0.5, 1]$ , we filter the original dataset  $D = \{\mathbf{x}_i, y_i, z'_i\}$  by keeping an instance  $i$  only if

it satisfies  $p(z_i = z'_i | \mathbf{x}_i) \geq \epsilon$ .

For well-calibrated classifiers like logistic regression, we expect to remove mostly mislabeled data points by thresholding at  $\epsilon$ . Making  $\epsilon$  vary between 0.5 and 1 allows us to modify the output of the preliminary study in order to obtain a sub-dataset with as many points correctly labeled as possible. Moreover, when the error of our preliminary classifier is symmetric, this process will also move the estimated correlation  $r(y, z')$  towards the true correlation  $r(y, z)$ .

With this smaller set of training instances, we run backdoor adjustment without modification. However, one important drawback of this method is that we remove instances from our training dataset. Depending on the quality of the preliminary classifier and the setting of  $\epsilon$ , only a small fraction of training instances may potentially remain. Thus, in the next section we consider an alternative approach that does not require discarding training instances.

**Correlation matching** While the above approach aims to reduce errors in  $z'$ , and as a side effect improves the estimate of  $r(y, z)$ , in this section we propose an approach that directly tries to improve the estimate of  $r(y, z)$  while also reducing errors in  $z$ . Let  $r' = r(y, z')$  be the observed correlation between  $y$  and  $z'$ , and let  $r = r(y, z)$  be the true (unobservable) correlation between  $y$  and  $z$  in the training data for  $y$ ,  $D = \{\mathbf{x}_i, y_i, z'_i\}$ . Our proposed approach builds on the insight of Francis, Coats, and Gibson (1999), who show that  $r'$  can be estimated from  $r$  using the variances of  $y$  and  $z$  as well as the variances of the errors in  $y$  and  $z$ :

$$r' = \sqrt{\frac{1}{(1 + \frac{V_{ey}}{V_y})(1 + \frac{V_{ez}}{V_z})}} \times r \quad (2)$$

where  $V_z$  is the variance of  $z$ , and  $V_{ez}$  is the variance of error on  $z$ , and analogously for  $V_y$ ,  $V_{ey}$ . Since in our setting  $y$  is observed, we can set  $V_{ey} = 0$  and solve for  $r$ :

$$r' = \sqrt{\frac{1}{1 + \frac{V_{ez}}{V_z}}} \times r \quad (3)$$

$$\Rightarrow r = r' \times \sqrt{1 + \frac{V_{ez}}{V_z}} \quad (4)$$

Thus, the factor by which  $r'$  underestimates  $r$  is proportional to the ratio of the variance of the error in  $z$  to the variance of  $z$ .

We can estimate the terms  $V_z$  and  $V_{ez}$  using cross-validation on the preliminary training data  $D_z = \{(\mathbf{x}_i, z_i)\}$ . Let  $z'_i$  be the value predicted by the preliminary classifier on instance  $\mathbf{x}_i \in D_z$ , where  $i$  is in the testing fold of one cross-validation split of the data. Let  $e_i^z = |z_i - z'_i|$  be the absolute error of  $z$  on instance  $i$ . Then, we can first compute the mean absolute error of  $z'_i$  as  $\mu_{ez} = \frac{1}{|D_z|} \sum_{z_i \in D_z} e_i^z$ . The estimated variance of the errors in  $z$  is then:

$$\hat{V}_{ez} = \frac{1}{|D_z|} \sum_{z_i \in D_z} (e_i^z - \mu_{ez})^2 \quad (5)$$

Since this variance in the error of  $z$  in turn affects the observed variance of  $z$ , we can then estimate

$$\hat{V}_z = V_{z'} - \hat{V}_{ez} \quad (6)$$

where  $V_{z'}$  is the variance of predictions  $z'$  in the target training data  $D$ .

Plugging the estimates of Equations 5 and 6 into Equation 4 enables us to estimate the true correlation between  $y$  and  $z$  in the target training data  $D$ . We will refer to this estimated correlation as  $\hat{r}$ .

As an example, consider a dataset  $D = \{(\mathbf{x}_i, y_i, z'_i)\}$ . The original correlation  $r(y, z') \equiv r'$  may be .5, but the true correlation  $r(y, z) \equiv r$  may be .8. Depending on the variances of  $z$  and its error, the estimated correlation may be  $\hat{r} = .75$ . The next step in the procedure is to optimize the assignment to  $z'$  to minimize the difference  $|r' - \hat{r}|$ . That is, we use  $\hat{r}$  as a soft constraint, and attempt to match that constraint by changing the assignments to  $z'$ .

Let  $\mathbf{Z}$  be the set of all possible assignments to  $z$  in the training set  $D$  (i.e., if  $z$  is a binary variable and  $|D| = n$ , then  $|\mathbf{Z}| = 2^n$ ). Let  $\mathbf{z}^j = \{z_1^j \dots z_n^j\} \in \mathbf{Z}$  be a vector of assignments to  $z$ , and let  $r'(\mathbf{z}^j)$  indicate the correlation  $r(\mathbf{z}^j, y)$ . Then our objective is to choose an assignment from  $\mathbf{Z}$  to minimize  $r'(\mathbf{z}^j) - \hat{r}$ , while still maximizing the probability of that assignment according to the preliminary classifier for  $z$ . We can write this objective as follows:

$$\mathbf{z}^* \leftarrow \arg \max_{\mathbf{z}^j \in \mathbf{Z}} \left( \frac{1}{n} \sum_{z_i^j \in \mathbf{z}^j} p(z_i = z_i^j | \mathbf{x}_i) \right) - |\hat{r} - r'(\mathbf{z}^j)| \quad (7)$$

Thus, we search for an optimal assignment  $\mathbf{z}^*$  that maximizes the average posterior of the predicted  $z$  value, while minimizing the difference between the estimated correlation  $\hat{r}$  and the observed correlation  $r'(\mathbf{z}^j)$ .

This optimization problem can be approached in several ways. We implement a greedy hill-climbing algorithm that iterates through the values in  $z'$  sorted by confidence and flips the value if it reduces  $|r - r'|$ . The steps are as follows:

1. Initialize  $\mathbf{z}^j$  to the most probable assignment according to  $p(z|\mathbf{x})$ .
2. Initialize  $\mathcal{I}$  to be all instances sorted in descending order of confidence  $p(z|\mathbf{x})$ .
3. While  $|\hat{r} - r'(\mathbf{z}^j)|$  is decreasing:
  - (a) Pop the next instance  $(\mathbf{x}_i, z_i^j, y_i)$  from  $\mathcal{I}$
  - (b) If flipping the label  $z_i^j$  reduces the error  $|\hat{r} - r'(\mathbf{z}^j)|$ , do so. Else, skip to the next instance.
4. Return the final  $\mathbf{z}^j$ .

For example, consider the case where  $r'(\mathbf{z}^j) < \hat{r}$ . If the instance popped in step 3(a) has labels  $(y_i = 1, z_i^j = 0)$ , then we know that flipping  $z_i$  to 1 would increase the correlation between  $y$  and  $z'$ . By considering flips in descending order of  $p(z|\mathbf{x})$ , we ensure that we first flip assignments that are likely to be incorrect. In the experiments below, we find that this approach often converges after a relatively small number of flips.

The advantages of this approach are that it not only produces assignments to  $z$  that better align with the expected correlation  $\hat{r}$ , but it also results in more accurate assignments to  $z$ . The latter is possible because we are using prior knowledge about the relationship between  $z$  and  $y$  to assign values

of  $z$  when the classifier is uncertain. As with the thresholding approach of the previous section, once the new assignments to  $z$  are found, back-door adjustment is run without modification.

## 4 Experiments

We conducted text classification experiments in which the relationship between the confounder  $z$  and the class variable  $y$  varies between the training and testing set. We consider the scenario in which we directly control the discrepancy between training and testing. Thus, we can determine how well a confounder has been controlled by measuring how robust the method performs across a range of discrepancy levels.

To sample train/test sets with different  $p(y|z)$  distributions, we assume we have labeled datasets  $D_{train}$ ,  $D_{test}$ , with elements  $\{(\mathbf{x}_i, y_i, z_i)\}$ , where  $y_i$  and  $z_i$  are binary variables. We introduce a bias parameter  $p(y = 1|z = 1) = b$ ; by definition,  $p(y = 0|z = 1) = 1 - b$ . For each experiment, we sample without replacement from each set  $D'_{train} \subseteq D_{train}$ ,  $D'_{test} \subseteq D_{test}$ . To simulate a change in  $p(y|z)$ , we use different bias terms for training and testing,  $b_{train}$ ,  $b_{test}$ . We thus sample according to the following constraints:  $p_{train}(y = 1|z = 1) = b_{train}$ ,  $p_{test}(y = 1|z = 1) = b_{test}$ ,  $p_{train}(Y) = p_{test}(Y)$ , and  $p_{train}(Z) = p_{test}(Z)$ .

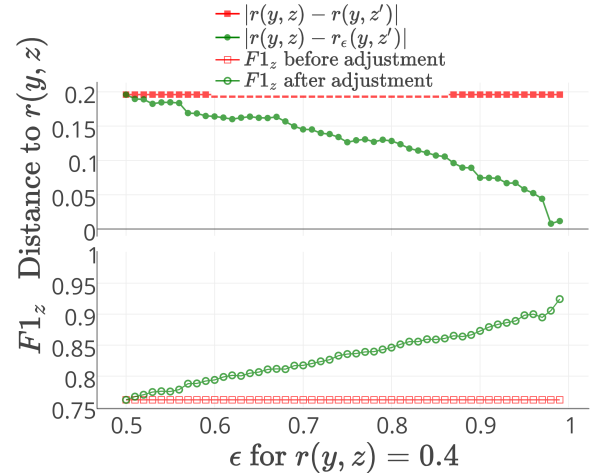
The last two constraints are to isolate the effect of changes to  $p(y|z)$ . Thus, we fix  $p(y)$  and  $p(z)$ , but vary  $p(y|z)$  from training to testing data. We emphasize that we do not alter any of the actual labels in the data; we merely sample instances to meet these constraints. In the rest of the paper, we note  $r_{train}(y, z)$  (respectively  $r_{test}(y, z)$ ) the correlation between  $y$  and  $z$  in the training set (resp. testing set). We also denote  $\delta_{yz} = r_{train}(y, z) - r_{test}(y, z)$ .

### 4.1 Datasets

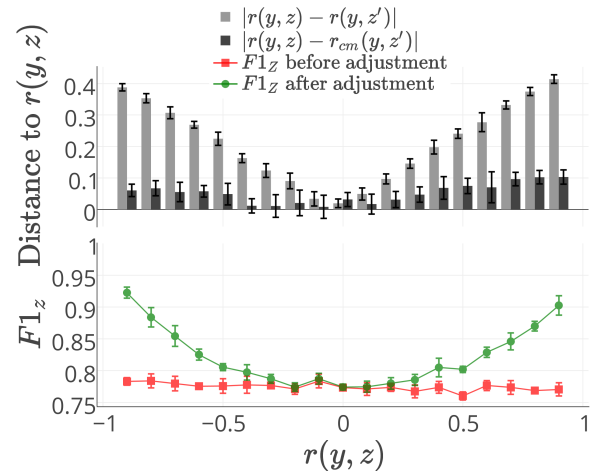
**Location / Gender** For our first dataset, we use the data from Landeiro and Culotta (2016), where the task is to predict the location of a Twitter user from their messages, with gender as a potential confounder. Thus,  $\mathbf{x}$  is a term vector,  $y$  is location, and  $z$  is gender. The data contain geolocated tweets from New York City (NYC) and Los Angeles (LA). There are 246,930 tweets for NYC and 218,945 for LA over a four-day period (June 15th to June 18th, 2015). Gender labels are derived by cross-referencing the user’s name (from the profile) with U.S. Census name data, removing ambiguous names. For each user, we have up to the most recent 3,200 tweets, which we represent each as a single binary unigram vector per user, using standard tokenization. Finally, we subsample this collection and keep the tweets from 6,000 users such that gender and location are uniformly distributed over the users.

**Smoker / Gender** In our second dataset, the task is to predict if a Twitter user is a smoker or not, with gender as a potential confounder. We start from approx. 3M tweets collected in January and February 2014 using cigarettes related keywords. We randomly pick 40K tweets for which we can identify the user’s gender using the Twitter screen name and the U.S. Census name data. We then manually annotate 4.5K

of these tweets on whether they show that a user is a smoker (yes) or a non-smoker (no) while discarding uncertain tweets (unknown). We use this data to train a classifier (F1 score = 0.84) to label the remaining 35.5K tweets on the smoker dimension. In order to avoid mislabeled tweets as much as possible, we only keep predictions with a confidence of at least 95%, yielding an additional 5.5K automatically labeled tweets. These 10K (4.5K manually annotated + 5.5K automatically annotated) tweets have been written by 9K unique users. For each of these users, we collect the most recent tweets (up to 200). Because some users set their profile to be private or because some users that existed in early 2014 have now deleted their account, we obtain at least 20 tweets for 4.6K users. Then we collect all the cigarettes related tweets published by a user in the first two months of 2014 and add them to our dataset. Finally, we remove users in order to build a balanced dataset on both annotated dimensions and eventually obtain a dataset of 4084 users.



(a) Effect of  $\epsilon$  thresholding on  $F1_z$  and distance to true correlation.



(b) Effect of correlation matching on  $F1_z$  and distance to true correlation.

Figure 2: Effect of correlation adjustment methods.

## 5 Results

We use the following notations to describe the results below:

- $\delta_{yz} = r_{train}(y, z) - r_{test}(y, z)$  is the discrepancy between the correlation of  $y$  and  $z$  in training versus testing.
- $r(y, z)$  (respectively  $r(y, z')$ ) is the true (resp. observed) correlation between  $y$  and  $z$ .
- $r(y, z'_\epsilon)$  (respectively  $r(y, z'_{cm})$ ) is  $r(y, z')$  after it has been adjusting using the  $\epsilon$  thresholding method (resp. the correlation matching method).
- $F1_z$  (respectively  $F1_y$ ) is the F1 score for a  $z$  (resp.  $y$ ) classifier, i.e. for the preliminary (resp. main) study.

### 5.1 Effects of correlation adjustments on $F1_z$

For this first part of our results, we obtain quasi-identical outcomes for both datasets. Therefore, we only present the results from the location/gender dataset.

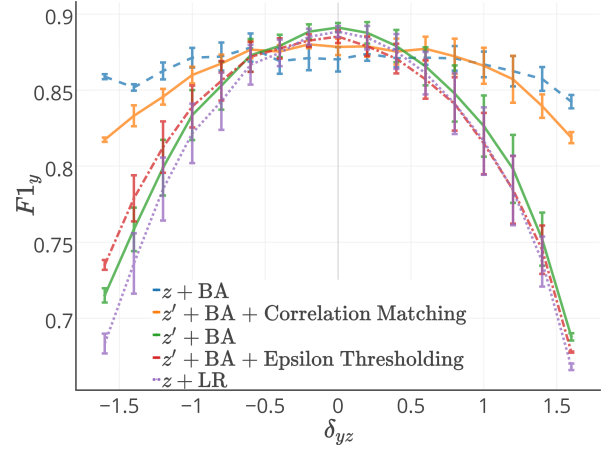
**$\epsilon$  thresholding method:** We make  $\epsilon$  vary between 0.5 and 0.95 and observe how this reduces the difference between  $r(y, z'_\epsilon)$  and  $r(y, z)$ . Figure 2(a) shows the result of one setting where  $r(y, z) = 0.4$ . The figure demonstrates that by increasing  $\epsilon$ ,  $r_\epsilon(y, z)$  gets closer to the true  $r(y, z)$ , and the performance of our external study is improved. This indicates that the classifier is well calibrated (since high confidence predictions are more likely to be correct). However, it takes a high value of  $\epsilon$  to get a correct approximation of the true association between  $y$  and  $z$ , meaning that we need to discard a large amount of data points from our preliminary study to approximate  $r(y, z)$ . For example, at  $\epsilon = .9$ , roughly half of the training instances remain.

**Correlation matching method:** For this method, we make the true correlation  $r(y, z)$  change between  $-0.8$  and  $0.8$  and we plot the results on Figure 2(b). We observe in the top plot that after adjustment, our estimate  $r_{cm}(y, z)$  is within 0.1 of the true correlation in the worst case against 0.4 without adjustment. This is a clear improvement in the correlation estimation. (For comparison, achieving a similarly accurate estimate using  $\epsilon$  thresholding requires removing 60% of 1500 instances.) We can also notice that the performance of our preliminary study greatly increases when we improve the estimation of  $r(y, z)$ , particularly when  $r(y, z)$  is high. For example, when  $r(y, z)$  is .8, the  $F1_z$  improves from .77 to .9, on average. Thus, correlation matching appears to both recover the true correlation while simultaneously improving the quality of the classifications of  $z$ .

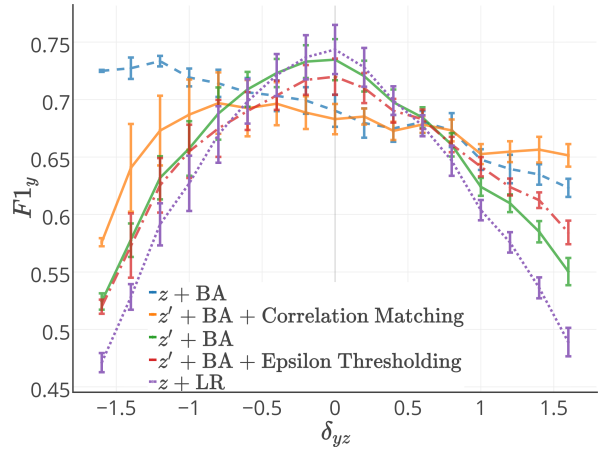
### 5.2 Effects of correlation adjustments on $F1_y$

#### Location / Gender

**Fixed  $F1_z = 0.784$ :** As our primary result, we report the  $F1_y$  obtained by different correlation adjustment methods across a range of shifts in the discrepancy between training and testing. For the Twitter dataset, the best performance we get in the preliminary study is  $F1_z = 0.784$ . We then compare testing  $F1_y$  as  $r_{train}(y, z)$  and  $r_{test}(y, z)$  vary. The results are shown in Figure 3(a). Without any adjustment, the performance we get is close to Logistic Regression. When using  $\epsilon$  thresholding, the performance is slightly improved



(a) Location/gender dataset



(b) Smoker/gender dataset

Figure 3:  $F1_y$  of the different adjustment methods when  $F1_z$  is fixed to its maximal value vs. logistic regression ( $z + LR$ ) and back-door adjustment ( $z + BA$ ) when  $z$  is observed.

$F1_z$	No adjustment	Corr. matching	$\epsilon$ thresh.
<b>0.784</b>	0.0640	0.0212	0.0610
<b>0.764</b>	0.0674	0.0313	0.0671
<b>0.702</b>	0.0677	0.0357	0.0803
<b>0.670</b>	0.0672	0.0345	0.0783
<b>0.645</b>	0.0705	0.0537	0.101
<b>0.557</b>	0.0715	0.124	0.0954
<b>0.519</b>	0.0709	0.0916	0.0941

Table 2: Standard deviation as a measure of robustness. The smaller the standard deviation, the more robust the model. The most robust model is shown in bold for each  $F1_z$  value.

in the extreme cases but only by a few points at most. However, when using the correlation matching method, we improve  $F1_y$  by 10 to 15 points in the most extreme cases. For comparison, the figure also shows the fully observed case



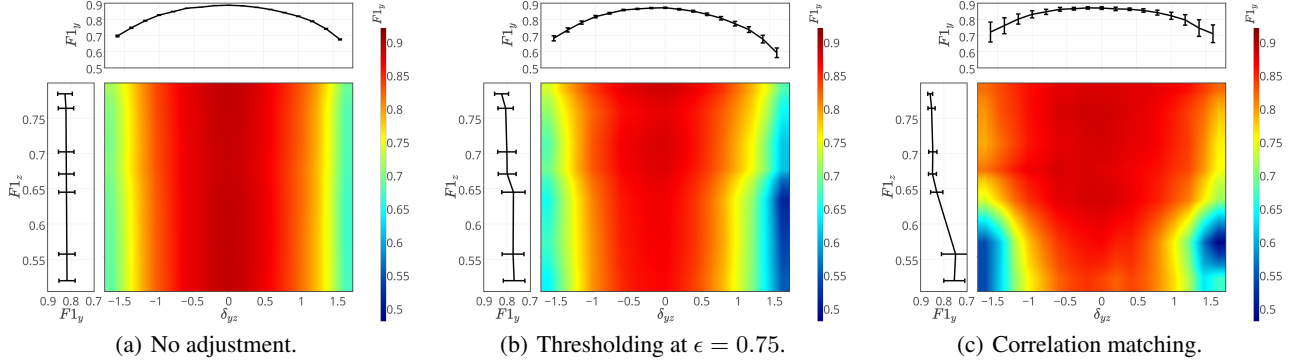


Figure 4: Experimental results for back-door adjustment with an *unobserved* confounding variable in the location/gender dataset.

( $z$ +BA), which uses back-door adjustment on the *true* values of  $z$ . We can see that correlation matching is comparable to the fully observed case, even with a 20% error rate on  $z$ . These results show that by getting a better estimate of the association between  $y$  and  $z$ , we can reduce attenuation bias and improve the robustness of our classifier, even though our observation of  $z$  is noisy.

**Variable  $F1_z$ :** We showed in the previous section that when we use our preliminary study with  $F1_z = 0.784$ , we can build a robust classifier using the correlation matching method combined with back-door adjustment. We also saw in Figure 1 that back-door adjustment when  $z$  is observed at training time is sensitive to noise in  $z$ . As a similar study, we want to see how sensitive the correlation adjustment methods are to the quality of  $F1_z$ . To do so, we increasingly add noise to the dataset used to train the preliminary classifier ( $D_z = \{\mathbf{x}_i, \mathbf{z}_i\}$ ) to make  $F1_z$  decrease. Because we want to visualize  $F1_y$  against two variables ( $F1_z$  and  $\delta_{yz}$ ), we visualize the results in a heatmap. In order to make the results clear to the reader, here are additional details to understand what is displayed on the heatmap: The x-axis of a heatmap is  $\delta_{yz}$  and the y-axis is  $F1_z$ . The line plot on the left of the heatmap shows  $F1_z$  given  $F1_y$  averaged over all possible values for  $\delta_{yz}$ . The error bars are the standard deviations of  $F1_y$ , indicating how sensitive the model is to variations of  $\delta_{yz}$ . Similarly, the scatter plot above the heatmap shows  $F1_y$  given  $\delta_{yz}$  averaged over all possible values for  $F1_z$ . The error bars are the standard deviations of  $F1_y$  for the matching  $\delta_{yz}$ .

Moreover, Table 2 displays the values of the standard deviations shown in the scatter plot at the left of each heatmap as a measure of robustness. Figure 4(a) shows the heatmap of results for back-door adjustment when we use the predictions of the preliminary study but none of the methods to fix the mislabeled values in  $z'$  are used. Figures 4(b) and 4(c) respectively show the heatmaps of results when we use  $\epsilon$  thresholding with  $\epsilon = 0.75$  and correlation matching. Similar to Figure 3(a),  $\epsilon$  thresholding only brings small improvement to no adjustment at all. Furthermore, when  $F1_z$  decreases, the correlation adjustment using  $\epsilon$  thresholding is performing worse than when we are not doing any corre-

lation adjustment as well as it is less robust. Clearly, the  $\epsilon$  thresholding method is more sensitive to the quality of the preliminary study than the other methods.

The correlation matching method (Figure 2(b)) does outperform the other methods in robustness and  $F1_y$  for most of the cases but when  $F1_z < 0.645$ , as we can see by the wider range of red values in Figure 4(c). In this latter case, it performs worse than the method without adjustment. This method is also sensitive to the quality of the preliminary study as we can see that the averaged  $F1_y$  decreases with  $F1_z$ . Let us remind one more time that we are considering here only preliminary studies with an  $F1_z$  of at most 0.784. Therefore,  $F1_z$  could be up to 22 points greater with a different dataset. This would hopefully lead to similar results than when  $F1_z = 0.784$  with correlation matching and better results in  $F1_y$  and robustness with  $\epsilon$  thresholding.

#### Smoker / Gender

**Fixed  $F1_z = 0.791$ :** Similarly to the previous experiment, we report  $F1_y$  while making  $\delta_{yz}$  vary as our primary result in Figure 3(b). We observe that predicting if a user smokes or not is a much more difficult task than our previous binary location prediction task, as the maximum yielded  $F1_y$  is around .75 when it was approximately .9 in the previous task. We also notice that the robustness of the back-door adjustment methods is not as good as for the location/gender dataset. The correlation matching method manages to perform closely to  $z + BA$  for  $\delta_{yz} \geq -0.75$  and outperforms all other methods for  $\delta_{yz} \geq 1$  but we also witness an accuracy drop on the left part of the plot. In addition to this drop, our two most robust methods ( $z + BA$  and correlation matching) are outperformed by approximately 5 points when there is no difference between the training correlation and the testing correlation (when  $\delta_{yz} = 0$ ).

**Variable  $F1_z$ :** When making  $F1_z$  vary with the smoker/gender dataset, we observe comparable outcomes as the ones displayed in the heatmaps of Figure 4 but with a lesser overall accuracy. As back-door adjustment was not performing as well as with the location/gender dataset in the fixed  $F1_z$  case, it logically also does not perform as well when  $F1_z$  varies. If we obtain a V-shaped heatmaps similar to Figures 4(b) and 4(c), the slope indicating that the

classifier's' performance deteriorates when  $F1_z$  decreases is steeper. This may show that our adjustments methods are more sensitive to noise in the confounding variable when the classification task is overall harder. We do not display the resulting heatmap for the smoker/gender experiment in this paper for brevity but we will make the dataset and the code to reproduce the results available online.

## 6 Conclusion

In this paper, we have proposed two methods of using back-door adjustment to control for an unobserved confounder. Using two real-life datasets extracted from Twitter, we have found that correlation matching on the predicted confounder associated with back-door adjustment can retrieve the underlying correlation  $r(y, z)$  and perform closely to back-door adjustment with an observed confounder. We also showed that  $\epsilon$  thresholding can be used to slightly improve the predictions compared to logistic regression. If  $\epsilon$  thresholding will not be able to adjust for the unobserved confounder  $z$  when  $F1_z < 0.75$ , we showed that correlation matching provides a way to adjust for an unobserved confounder and outperform plain back-door adjustment as long as  $F1_z > 0.65$ . In future work, we will consider hybrid methods that combine  $\epsilon$  thresholding and correlation matching to increase robustness as  $F1_z$  decreases.

## References

- [Angwin et al. 2016] Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. *ProPublica* 23.
- [Bareinboim, Tian, and Pearl 2014] Bareinboim, E.; Tian, J.; and Pearl, J. 2014. Recovering from selection bias in causal and statistical inference. In *Proceedings of The Twenty-Eighth Conference on Artificial Intelligence (CE Brodley and P. Stone, eds.)*. AAAI Press, Menlo Park, CA.
- [Bickel, Brückner, and Scheffer 2009] Bickel, S.; Brückner, M.; and Scheffer, T. 2009. Discriminative learning under covariate shift. *Journal of Machine Learning Research* 10(Sep):2137–2155.
- [Chesher 1991] Chesher, A. 1991. The effect of measurement error. *Biometrika* 78(3):451–462.
- [Cunha, Weber, and Pappa 2017] Cunha, T. O.; Weber, I.; and Pappa, G. L. 2017. A warm welcome matters! the link between social feedback and weight loss in/t/loseit. *arXiv preprint arXiv:1701.05225*.
- [De Choudhury et al. 2016] De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2098–2110. ACM.
- [De Choudhury, Counts, and Horvitz 2013] De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3267–3276. ACM.
- [Francis, Coats, and Gibson 1999] Francis, D. P.; Coats, A. J.; and Gibson, D. G. 1999. How high can a correlation coefficient be? effects of limited reproducibility of common cardiological measures. *International journal of cardiology* 69(2):185–189.
- [Fukuchi, Sakuma, and Kamishima 2013] Fukuchi, K.; Sakuma, J.; and Kamishima, T. 2013. Prediction with model-based neutrality. In *Machine Learning and Knowledge Discovery in Databases*. Springer. 499–514.
- [Hajian and Domingo-Ferrer 2013] Hajian, S., and Domingo-Ferrer, J. 2013. A methodology for direct and indirect discrimination prevention in data mining. *Knowledge and Data Engineering, IEEE Transactions on* 25(7):1445–1459.
- [Hand and Henley 1997] Hand, D. J., and Henley, W. E. 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160(3):523–541.
- [King, Keohane, and Verba 1994] King, G.; Keohane, R. O.; and Verba, S. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press.
- [Koratana et al. 2016] Koratana, A.; Dredze, M.; Chisolm, M. S.; Johnson, M. W.; and Paul, M. J. 2016. Studying anonymous health issues and substance use on college campuses with yik yak. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.
- [Kuroki and Pearl 2014] Kuroki, M., and Pearl, J. 2014. Measurement bias and effect restoration in causal inference. *Biometrika* 101(2):423–437.
- [Landeiro and Culotta 2016] Landeiro, V., and Culotta, A. 2016. Robust text classification in the presence of confounding bias. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- [Lazer et al. 2009] Lazer, D.; Pentland, A. S.; Adamic, L.; Aral, S.; Barabasi, A. L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; et al. 2009. Life in the network: the coming age of computational social science. *Science (New York, NY)* 323(5915):721.
- [Miller 2015] Miller, C. C. 2015. Can an algorithm hire better than a human? *The New York Times* 25.
- [Monahan and Skeem 2016] Monahan, J., and Skeem, J. L. 2016. Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology* 12:489–513.
- [Paul and Dredze 2011] Paul, M. J., and Dredze, M. 2011. You are what you tweet: Analyzing twitter for public health. *ICWSM* 20:265–272.
- [Pearl 2003] Pearl, J. 2003. Causality: models, reasoning and inference. *Econometric Theory* 19:675–685.
- [Pedreshi, Ruggieri, and Turini 2008] Pedreshi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 560–568. ACM.
- [Pennacchiotti and Popescu 2011] Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to twitter user classification. *ICWSM* 11(1):281–288.
- [Quionero-Candela et al. 2009] Quionero-Candela, J.;



- Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. D. 2009. *Dataset shift in machine learning*. The MIT Press.
- [Sugiyama, Krauledat, and Müller 2007] Sugiyama, M.; Krauledat, M.; and Müller, K.-R. 2007. Covariate shift adaptation by importance weighted cross validation. *The Journal of Machine Learning Research* 8:985–1005.
- [Tsymbal 2004] Tsymbal, A. 2004. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin* 106.
- [Widmer and Kubat 1996] Widmer, G., and Kubat, M. 1996. Learning in the presence of concept drift and hidden contexts. *Machine learning* 23(1):69–101.
- [Zadrozny 2004] Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, 114. ACM.
- [Zemel et al. 2013] Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 325–333.